Sociotechnical Safety Evaluation of Generative AI Systems

Year: 2023 | Citations: 179 | Authors: Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, L. Hendricks

Abstract

Generative AI systems produce a range of risks. To ensure the safety of generative AI systems, these risks must be evaluated. In this paper, we make two main contributions toward establishing such evaluations. First, we propose a three-layered framework that takes a structured, sociotechnical approach to evaluating these risks. This framework encompasses capability evaluations, which are the main current approach to safety evaluation. It then reaches further by building on system safety principles, particularly the insight that context determines whether a given capability may cause harm. To account for relevant context, our framework adds human interaction and systemic impacts as additional layers of evaluation. Second, we survey the current state of safety evaluation of generative AI systems and create a repository of existing evaluations. Three salient evaluation gaps emerge from this analysis. We propose ways forward to closing these gaps, outlining practical steps as well as roles and responsibilities for different actors. Sociotechnical safety evaluation is a tractable approach to the robust and comprehensive safety evaluation of generative AI systems.